DOCUMENT RESUME

ED 210 276 TM 810 728

AUTHOR McCormick, Douglas; And Others
TITLE Empirical Identification of Hierarchies.
SPONS AGENCY Department of Justice, Washington, D.C. National Inst. of Justice.
PUB DATE Apr 81
GRANT 79-NI-AX-0065
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981). Small print in figures.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Cluster Analysis; Factor Analysis; *Research Methodology; *Statistics
IDENTIFIERS Empirical Analysis; *Hierarchical Cluster Analysis; *Matrix Operations; Order Analysis

ABSTRACT

Outlining a cluster procedure which maximizes specific criteria while building scales from binary measures using a sequential, agglomerative, overlapping, non-hierarchic method results in indices giving truer results than exploratory facotr analyses or multidimensional scaling. In a series of eleven figures, patterns within cluster histories reveal the structure of the data. If true clusters exist in the data, one way they reveal themselves is by a sharp drop in the index values as an item outside the true cluster is added. In spatial terms, this represents a "moat" surrounding the cluster: a low region of density between regions of higher density which are the clusters themselves. A series of analyses were conducted using artificial data which had a known cluster structure. The Birnbaum test model was used to produce unidimensional scales of three sizes, which were combined with six outliners to make the raw data for analysis. Means, variances, and distribution shapes were varied for the Birnbaum parameters of difficulty, ability and discrimination. (Author/CP)

Empirical Identification of Hierarchies

Douglas McCormick

Robert Cudeck

Norman Cliff

A paper presented to the American Educational Research Association at the annual meetings in Los Angeles, April 13 - 17, 1981.*

*Funding for this research was provided by the National Institute of Justice Grant #79-NI-AX-0065.

Cliff and Reynolds (paper presented in this session) have suggested a class of indices which avoids the problems encountered when correlations are used with binary data and promises to increase validity when used to construct scales or hierarchies.

The new indices don't correspond functionally with variances the way correlations do. This means that although test items or other binary data can be factor analysed using the new indices, one is no longer accounting for the maximum amount of variance in each successive factor and it's unclear that any other property is explicitly maximized instead.

Exploratory factor analyses could still be done and so could multidimensional scaling, although multidimensional scaling has recieved even less attention than factor analysis as a method for finding scales in binary data. I'm going to talk instead about a clustering procedure which allows one to maximize specific criteria while building scales from binary measures.

In the language of cluster analysis(Sneath and Sokal,1973), this is a sequential,agglomerative,overlapping, non-hierarchic method.

In Table 1 is an example of what our computer program has done with a set of artificial data.

The first matrix is a matrix of the indices themselves. The data consist of eight items. The first six belong together in a scale and the last two are only randomly associated with the scale or each other.

The program begins a cluster or scale with each of the eight items. To each cluster is added the item with the highest average index with items already in the cluster. The record of the clustering process is kept in two matrices. The first records the average index as each item is added. The first entry in column one from the record of 0's, shows that the cluster began with item one. Below are the index values as each new item joined cluster one. The "item history" matrix records the item numbers in the order in which they were added.

Patterns within these cluster histories reveal the structure of the data. If true clusters exist in the data, one way they reveal themselves is by a sharp drop in the index values as an item outside the true cluster is added. In spatial terms this represents a "moat" surrounding the cluster; a region of low density between regions of higher density which are the clusters themselves.

The next matrix represents the membership in each cluster which results if we stop each cluster history where the largest gap occurs. Here each column still represents the cluster began with each item, but the rows here also represent the items in order, so the 1's in column one mean the first cluster contains items 1 through 6.

You can see from the rectangular block of 1's that items 1 through 6 in fact form a cluster. Items 7 and 8 join the cluster only when they start their own and can

furthermore be identified as outliers by the low index values with which their clusters began. This low initial value then rises as the cluster moves to a region of higher density. This is the opposite to the pattern of a legitimate cluster where the values decrease.

In situations where there is no moat but only gradual thinning and rethickening, an alternative to looking for a gap is to set a cutoff value below which items cannot enter a cluster. The final matrix in Table 1 shows the result of using a cutting score of .10 to define the clusters in our artificial data. The correct items all form the cluster with items 1 through 6 and 7 and 8 are isolated in a perfect solution.

Before analyzing real data, a series of analyses were conducted using artificial data which had a known cluster structure. The Birnbaum test model was used to produce unidimensional scales of three sizes. These three scales were then combined with six outliers to make the raw data for analysis. Means, variances and distribution shapes were varied for the Birnbaum parameters of difficulty, ability and discrimination.

Figure one represents a cleaner than average, but not atypical solution. The vertical lines separate the three clusters and the six outliers which are on the right side of the diagram. The outliers again are identifiable because they are included only in the clusters they begin themselves and each one appears as an extra member added to one of the legitimate clusters. The "block diagonal" pattern is characteristic of a good cluster solution. We were able to obtain this type of solution with artificial data of surprisingly low quality.

When we began analyzing real data, block diagonals were harder to come by.

One of the first datasets we analyzed was composed of vocabulary test data. Items had been taken from several distinct content areas, medicine, transportation, finance and others as well. What we hoped for were separate clusters in each domain.

What we got with the largest gap rule was one big scale and inconsistent smaller scales. The rows and columns in Figure 2 have been reordered to show the pattern of the solution. The solution shown in Figure 3, which involves almost all of the clusters and nearly identical sets of items, was obtained by using a cutoff value of .47. The history of item numbers added to each cluster is shown in Figure 4.

There are two characteristics of the data that deserve comment. First is the fact that the same items are almost always the last to enter every cluster. These are the items which have lines drawn through them. Lines are also drawn through the clusters they began. For instance, item 21 is the last item to enter every cluster except the one which it began. It therefore seemed sensible to remove item 21 as an outlier.

The second noteworthy feature of the solution is that the same items seem to be among the first to be added to every cluster. The items which are circled in Figure 4 are the same eight items in every cluster. Most of the items act like outliers and add this set of items to their own cluster quickly after it has begun.

Factor analysis of all the items also suggests that they are largely univocal. Although the factor results are distorted by correlations calculated for items of unequal difficulty, the eight items which are circled are among those with the highest loadings on the first principle component.

The first dataset which showed more than one cluster was an adjective checklist of 50 or so items. Social workers had checked each adjective which applied to each of 256 mothers of delinquent boys.

Analysis of the adjectives using the largest gap to define the cluster boundaries produced one fairly well defined cluster and a lot of smaller clusters, some of which are related. The ordered solution is shown in Figure 5.

Clusters with overlapping membership often differ in the same way that items differ in a Guttman scale. Smaller clusters lack many of the items in the larger clusters because they ended sooner. As a way to identify these relationships we found it was possible to reanalyze the membership matrix that resulted from the first clustering.

The results from this second order clustering of the adjective checklist produced the membership matrix shown in Figure 6 which we have been calling a "supercluster" membership matrix.

Each row in this matrix represents one of the clusters which was a column in the previous membership matrix. Each column is a supercluster. There are eight unique patterns of superclusters shown here.

When we worked backwards and unique computed the relationship between individual items and superclusters, we obtained a matrix of loadings much like factor loadings.

With the exception of a few adjectives dealing with stubbornness and aggression, within the eight patterns there were really only two clusters. One consisted of all the positive qualities and one of all the negative qualities. The gap which separated the two groups was typically between positive and negative indices and close to zero. Although this is a legitimate two cluster structure, by using a cutoff near .5 it was possible to obtain a set of "superclusters" with a more interesting structure. This is shown in Figure 7. Besides being a cleaner solution, this strategy showed substructure in both the negative and positive items which had been concelled when the index was allowed to come near zero.

The final set of data looked at is a very large file of criminal offenses committed by a birth cohort of 28,000 young men born in Copenhagen. This data is unique, both for its completeness and its accuracy. It allowed us to test

the nction that delinquency is a unitary phenomenon against the competing hypothesis that delinquency can be subdivided into distinct criminal specialties.

Figures 8 and 9 show the first order cluster histories for a selected sample of 56 crimes. I have circled 10 of the crimes and you can see they are near the top of most clusters.

With the exception of a cluster for items 17,50,51 and 52 it appears delinquency has something in common with taking a vocabulary test.

If a cutoff point is chosen which preserves the cluster of these exceptions, the membership matrix which results is that displayed in Figure 10.

In the second order analysis, shown in Figure 11, the pattern is even clearer. The small cluster in the corner consists of traffic offenses; speeding, illegally overtaking, failure to yeild and item 17 which was negligent homicide.

The items most important in the larger cluster, those which were circled, include burglary, forgery, fraud, robbery and recieving stolen goods which are apparently core crimes in a scale or factor of general delinquency.

That's all the data we have looked at so far. We have tried to make the procedure more automatic, to reduce the influence of the researcher on the results and may make progress in that direction. As it operates now the program appears to be capable of uncovering a variety of data structures as I hope these examples demonstrate.

An Artificial Example with Eight Variables

1.00	.22	.14	.31	.17	.24	-.03	.01	
.22	1.00	.22	.13	.16	.13	-.05	.06	
.14	.22	1.00	.22	.17	.22	-.02	-.07	
.31	.13	.22	1.00	.09	.17	.01	-.09	
.17	.16	.17	.09	1.00	.13	.00	-.12	Q
.24	.13	.22	.17	.13	1.00	-.26	-.14	
-.03	-.05	-.02	.01	.00	-.26	1.00	-.05	
.01	.06	-.07	-.09	-.12	-.14	-.05	1.00	
1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	
.31	.22	.22	.31	.17	.24	.01	.06	
.21	.23	.21	.21	.20	.25	.14	.09	
.19	.19	.21	.19	.18	.19	.12	.12	Record of Q's
.17	.19	.17	.17	.18	.17	.14	.12	
.15	.15	.15	.15	.17	.15	.12	.13	
-.05	-.05	-.05	-.05	-.05	-.05	.12	.10	
-.06	-.06	0.06	0.06	-.06	-.06	-.06	-.05	
1	2	3	4	5	6	7	8	
4	1	4	1	1	1	4	2	
6	4	1	6	4	4	1	1	
3	3	6	3	6	3	3	4	
2	6	2	2	3	2	2	3	Item History
5	5	5	5	2	5	5	6	
7	7	7	7	7	7	6	5	
8	8	8	8	8	8	8	7	
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	Membership by
1	1	1	1	1	1	1	1	Largest Cap
1	1	1	1	1	1	1	1	
0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	1	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	
1	1	1	1	1	1	0	0	Membership by > .10
1	1	1	1	1	1	0	0	
0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	1	

Table 1.

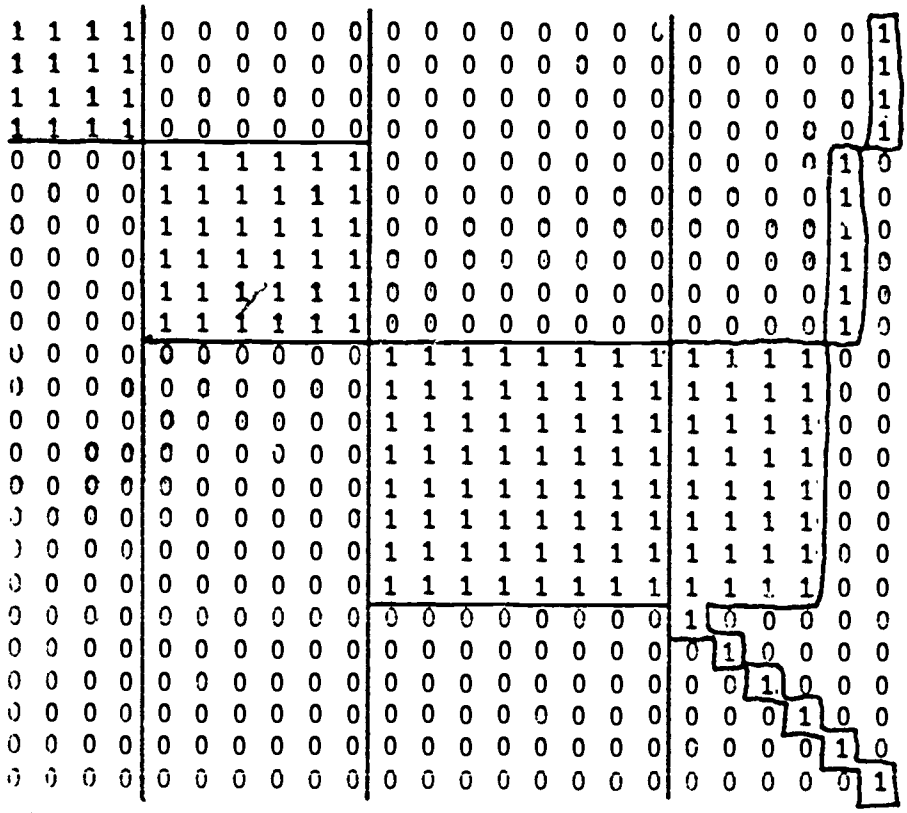


Figure 1.

2

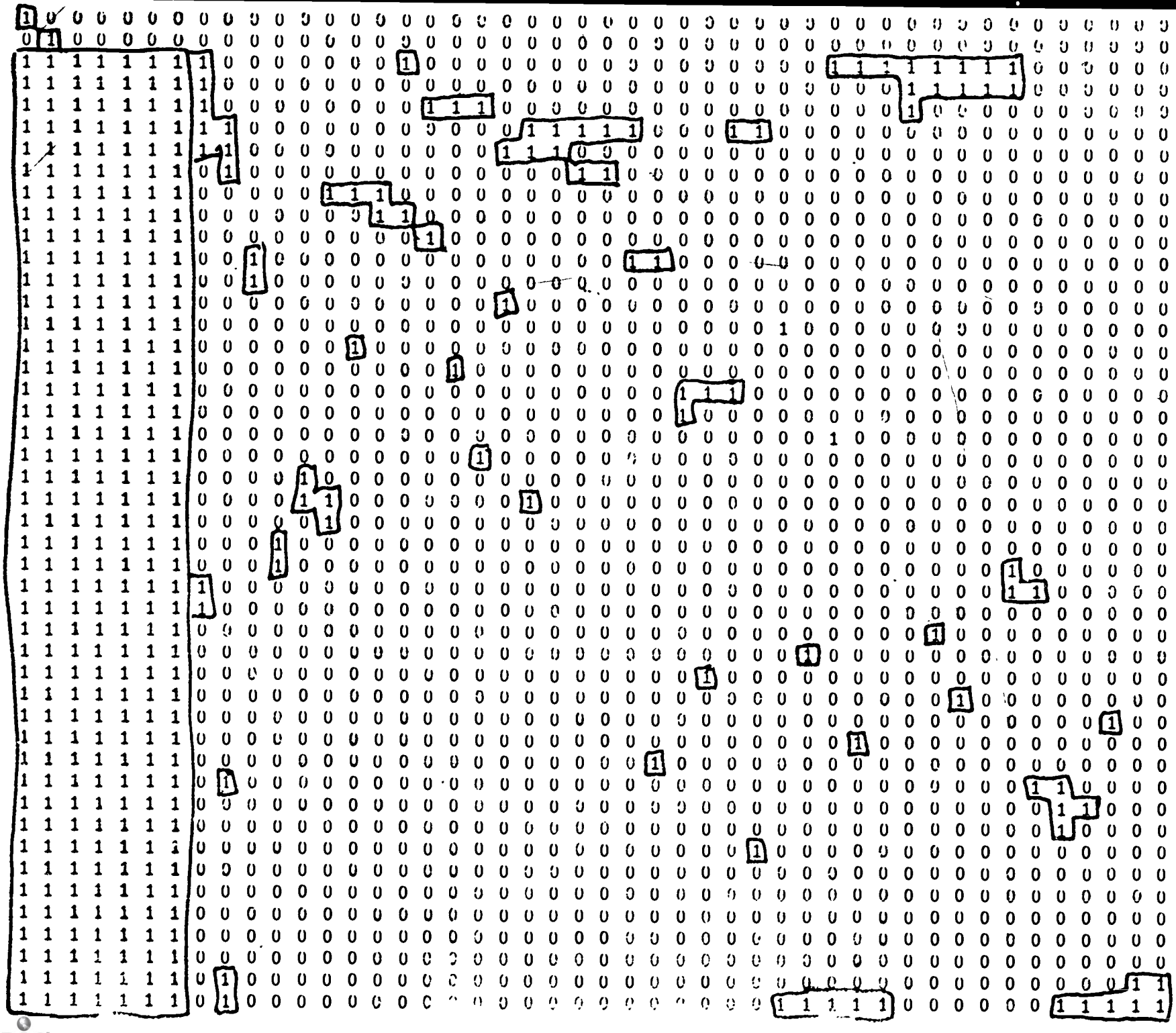


Figure 2.

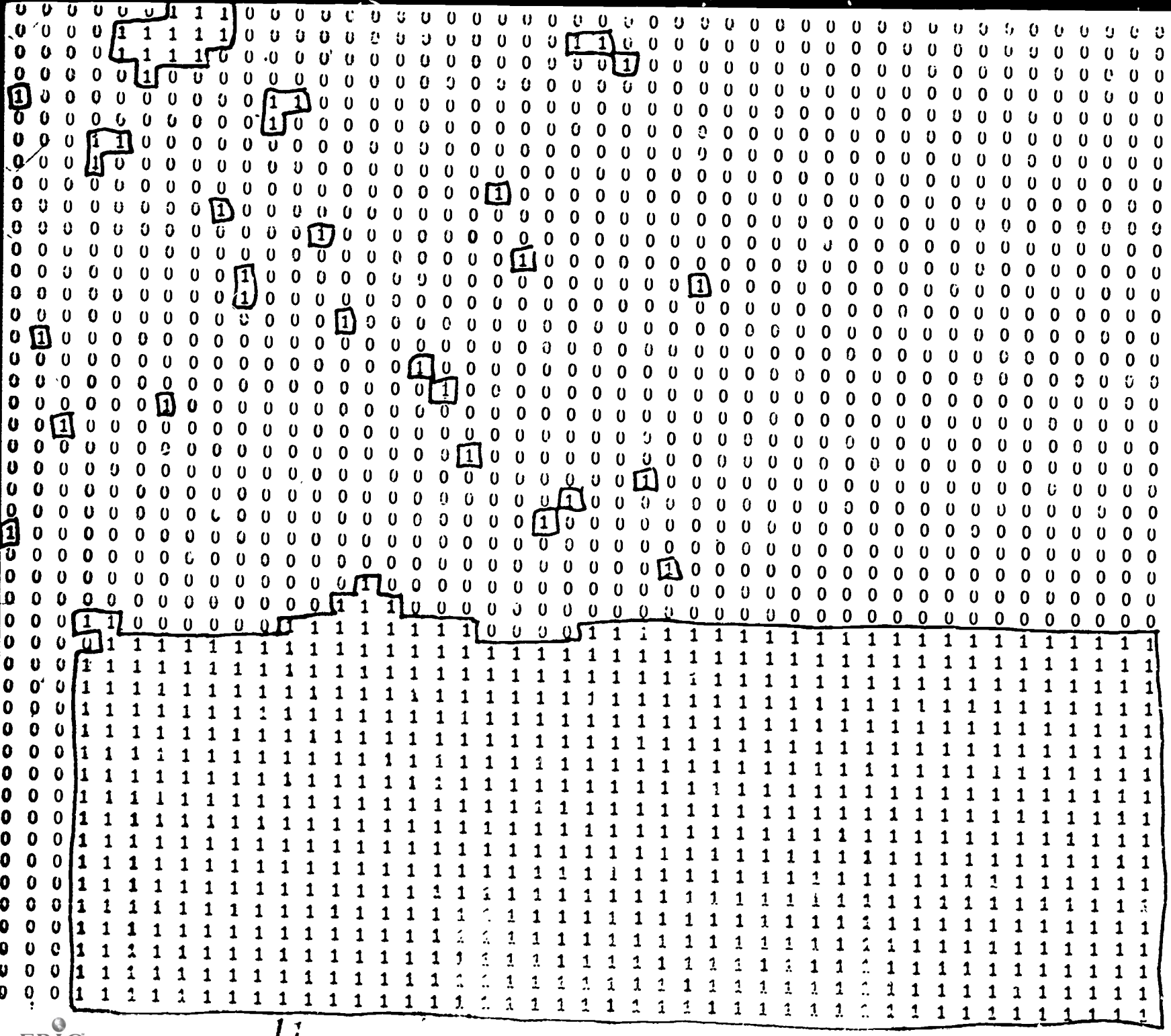
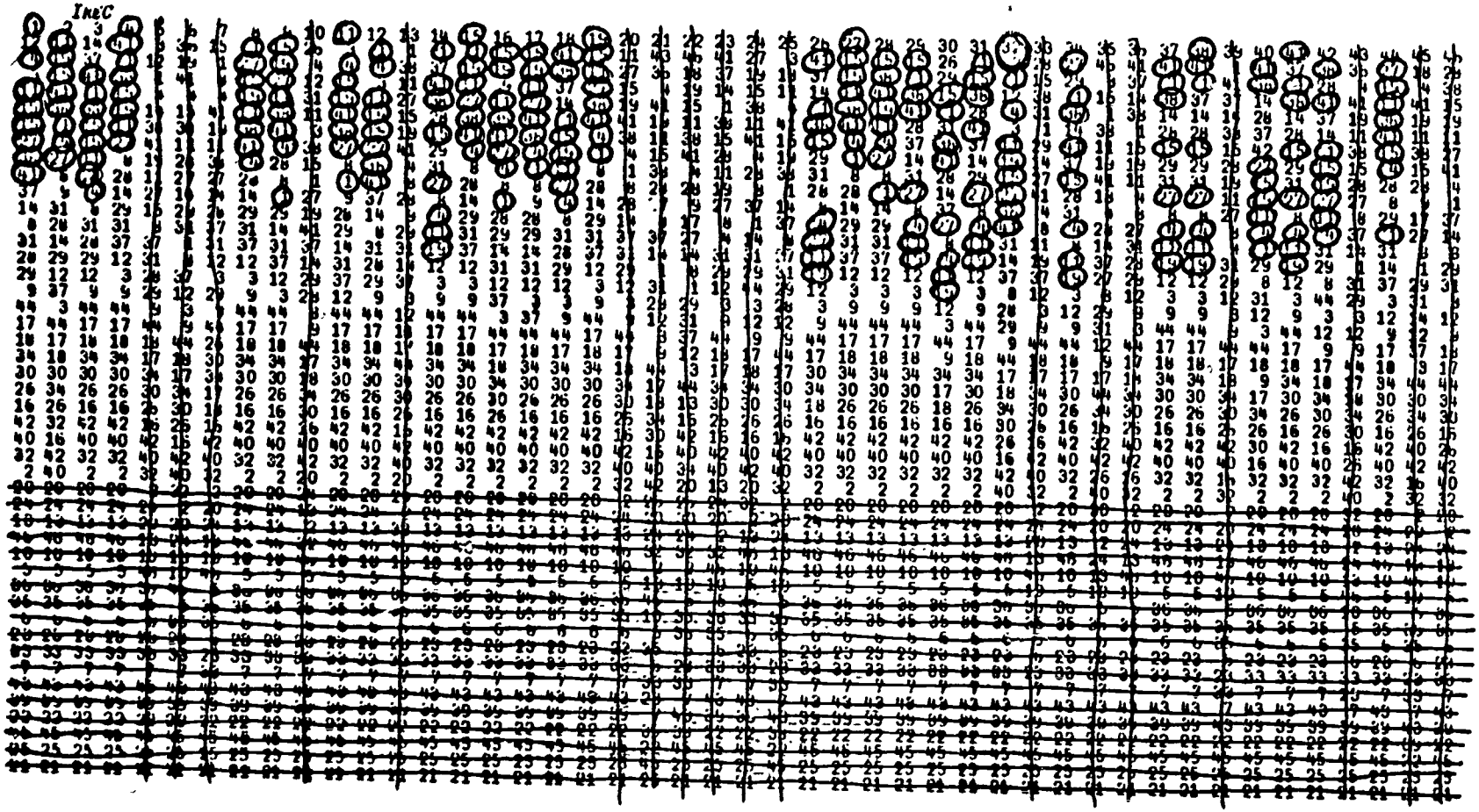


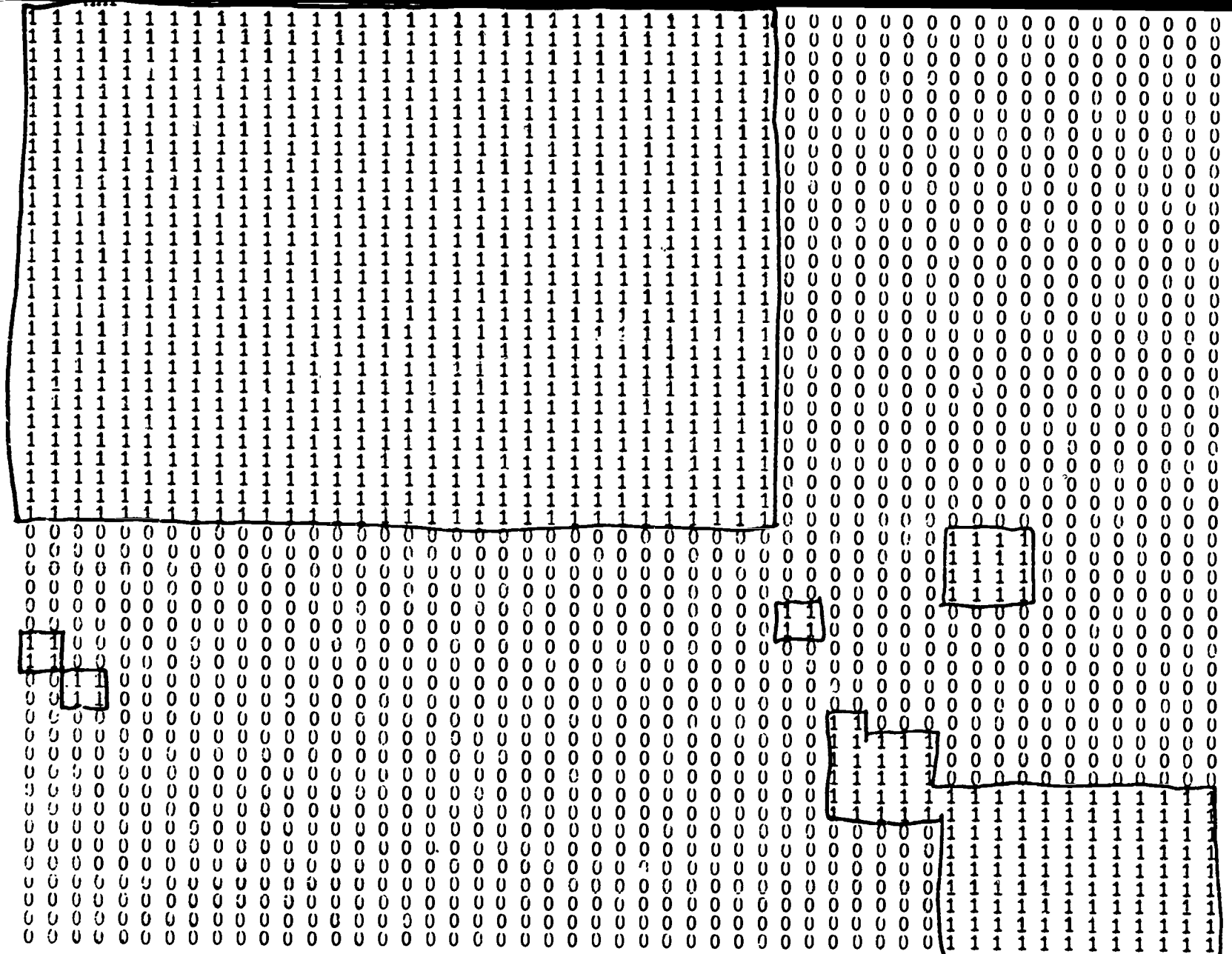
Figure 3.



1 4 11 15 19 27 38 41

Figure 4.

Figure 5.



8 7 1 6 4 5 2 3

17

(7)

18

Figure 6.

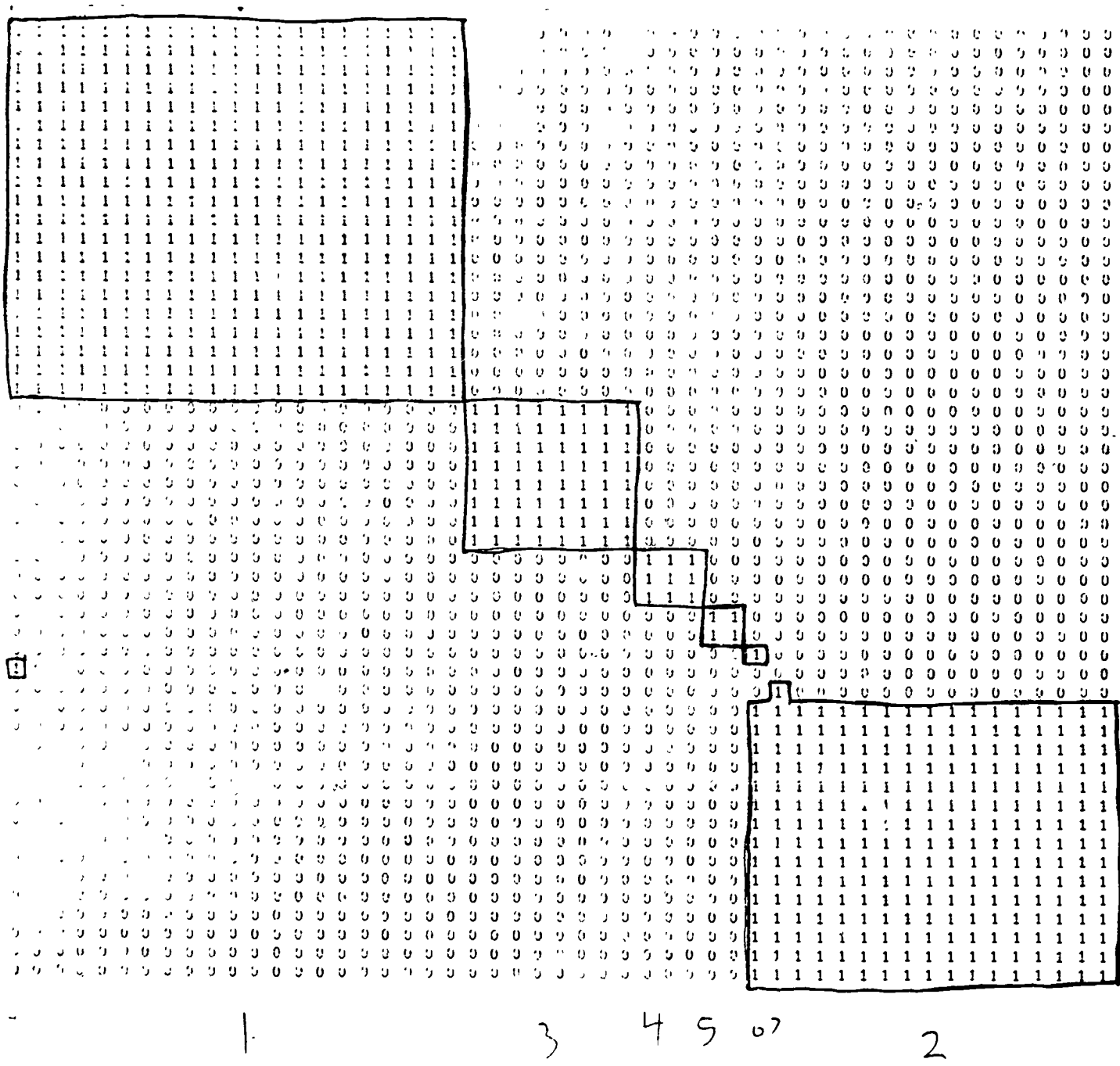


Figure 7.

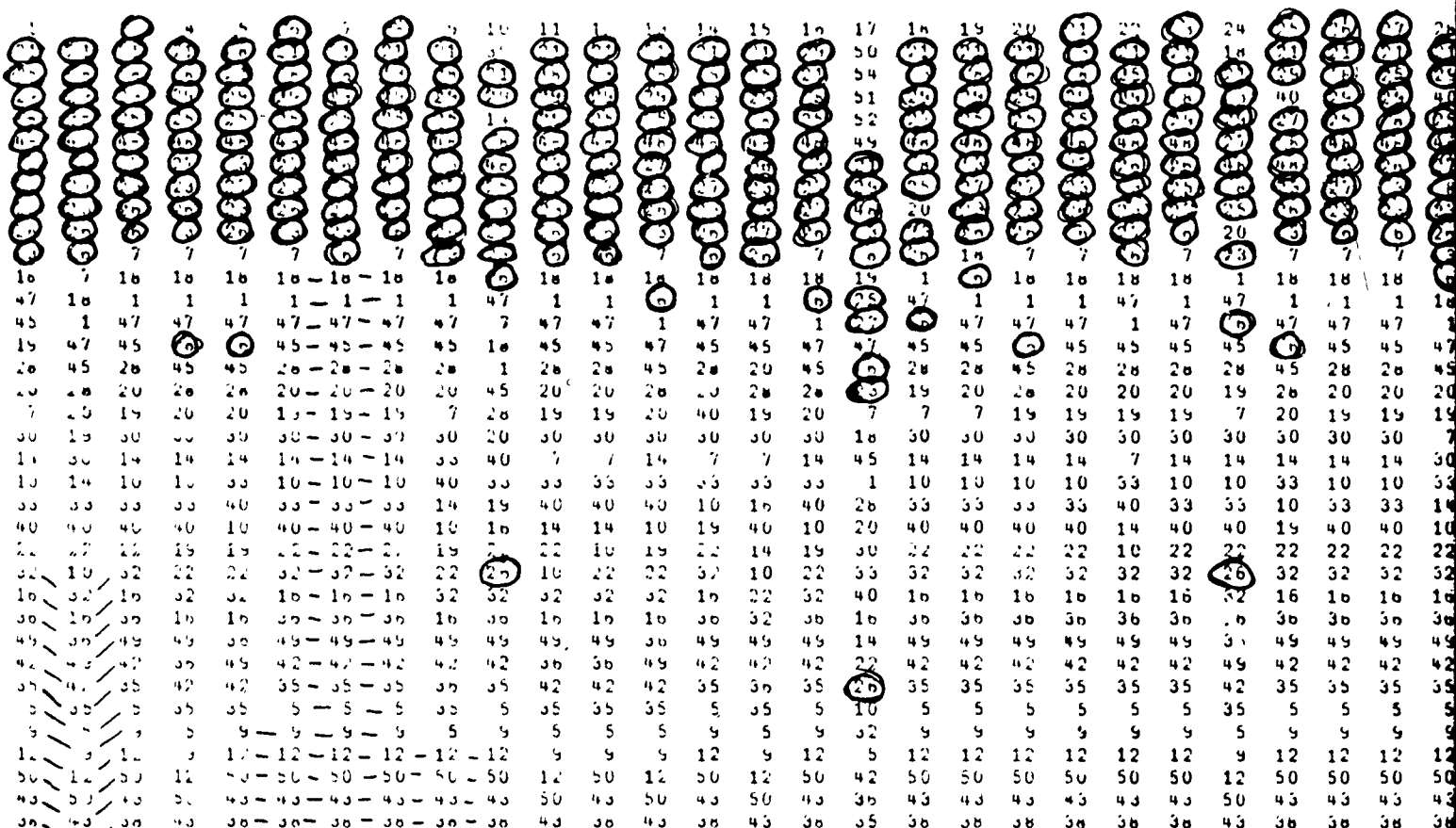
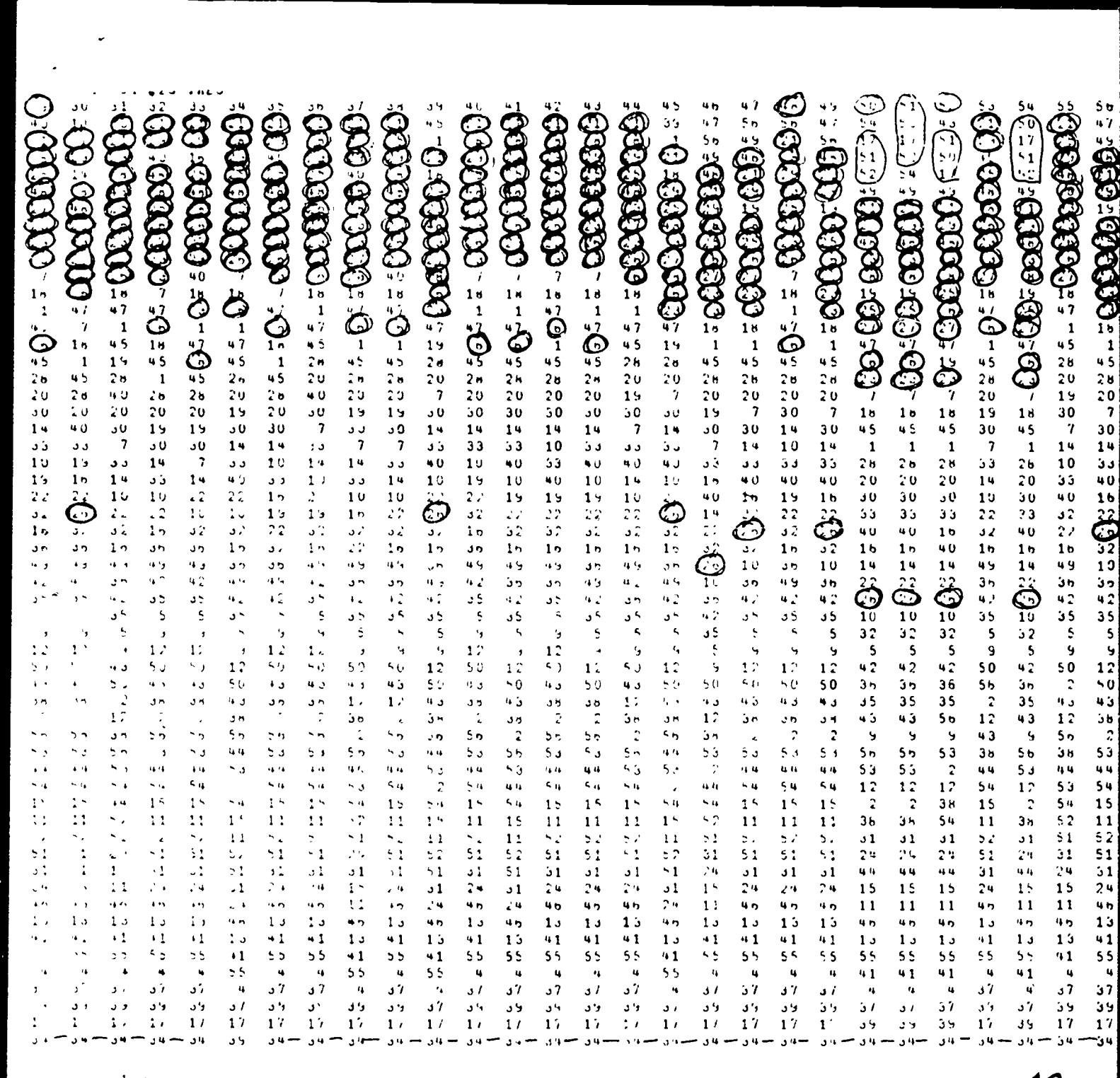
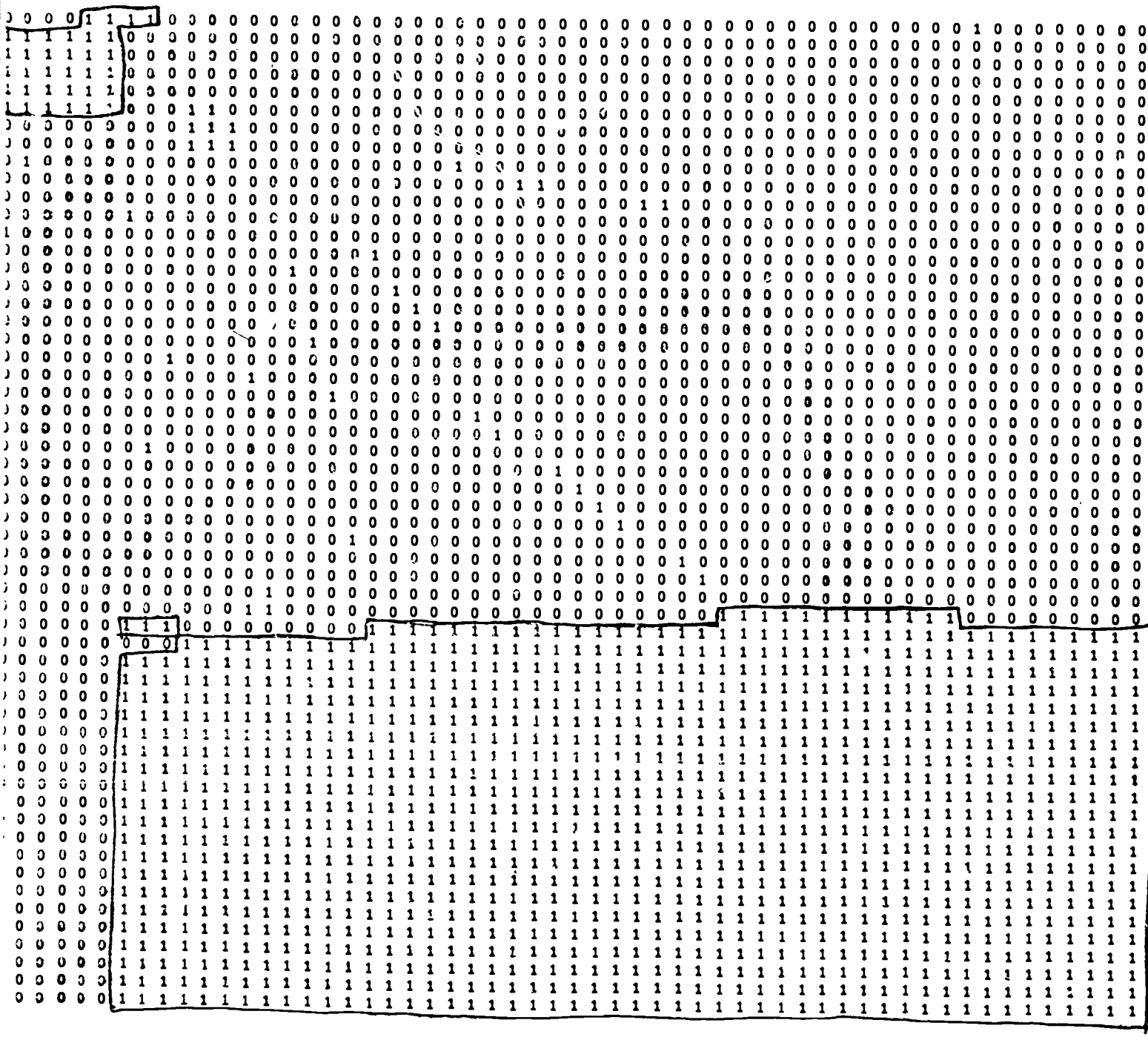


Figure 8.



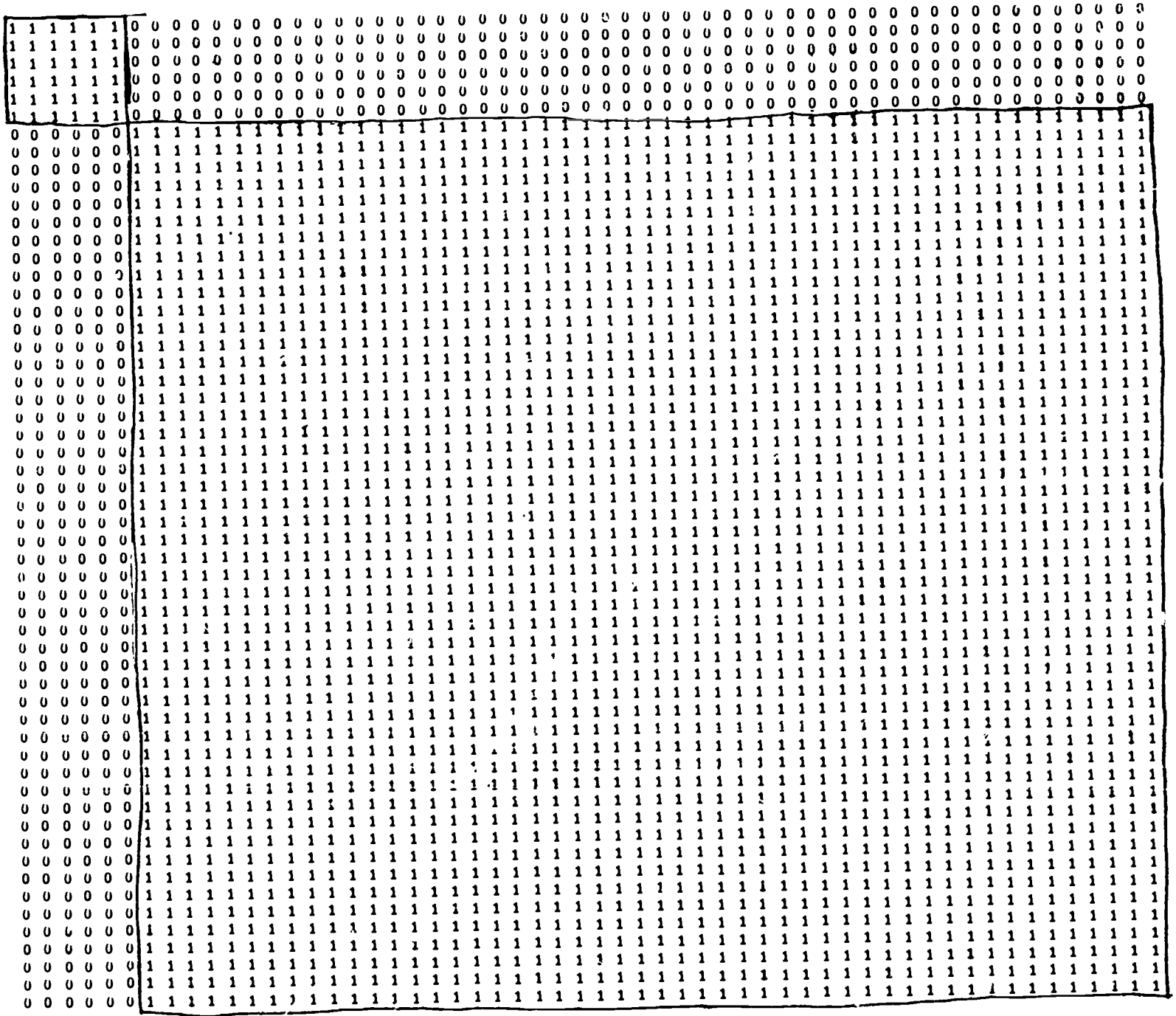
10

Figure 9.



11

Figure 10.



ICOPY 300024 SAMPB:SPC01 IREC QREC
SAVED 16:06:22 09/16/80

Figure 11.